Face Painting: querying art with photos

Elliot J. Crowley
elliot@robots.ox.ac.uk
Omkar M. Parkhi
omkar@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

Abstract

We study the problem of matching photos of a person to paintings of that person, in order to retrieve similar paintings given a query photo. This is challenging as paintings span many media (oil, ink, watercolor) and can vary tremendously in style (caricature, pop art, minimalist).

We make the following contributions: (i) we show that, depending on the face representation used, performance can be improved substantially by learning – either by a linear projection matrix common across identities, or by a per-identity classifier. We compare Fisher Vector and Convolutional Neural Network representations for this task; (ii) we introduce new datasets for learning and evaluating this problem; (iii) we also consider the reverse problem of retrieving photos from a large corpus given a painting; and finally, (iv) using the learnt descriptors, we show that, given a photo of a person, we are able to find their doppelgänger in a large dataset of oil paintings, and how this result can be varied by modifying attributes (e.g. frowning, old looking).

1 Introduction

Is there a painting of you out there? Probably not. But there may be one which looks just like you, as one man found out to his astonishment [6]. This raises the question of how to find such a painting (in a very large corpus) given a photo of a person's face. Of course, the extent to which a person in a photo resembles a different person in a painting is subjective, and very difficult to quantify. So, instead, we consider the question: given *photographs* of a person, can we retrieve *paintings* of that same person (in a large corpus)? The advantage of this question is that it is quantifiable, and so we can define a loss function and use tools from machine learning to improve performance. Also, armed with the developed methods we should then be able to find different, but similar looking, people in paintings, starting from a photo.

Initially, one might be skeptical over whether retrieving paintings of a person starting from a photo is achievable. Photographs and paintings have very different low level statistics and to make matters worse, painted portraits are prone to large variations in style: politicians are often highly caricatured, Hollywood icons of the past frequently get the Andy Warhol treatment and are transformed into pop art. This problem is essentially one of domain adaptation [15, 23, 32] from faces in photos to those in paintings; learning how to overcome both the low-level and stylistic differences.

Pages 65.1-65.13
DOI: https://dx.doi.org/10.5244/C.29.65

To investigate how successfully we can use face photos to retrieve paintings, we require a large corpus for which there are both photos and paintings of the same person. To this end we use photos of celebrities and public figures to retrieve paintings from two distinct datasets: (i) paintings from the National Portrait Gallery, which are largely photo-realistic in nature, and (ii) paintings produced by the public crawled from the DeviantArt website [3], which are much more varied in style. The contrast between these datasets allows us to observe what effect large variations in style have on retrieval. Figure 1 shows samples from the datasets, which are described in more detail in section 3.

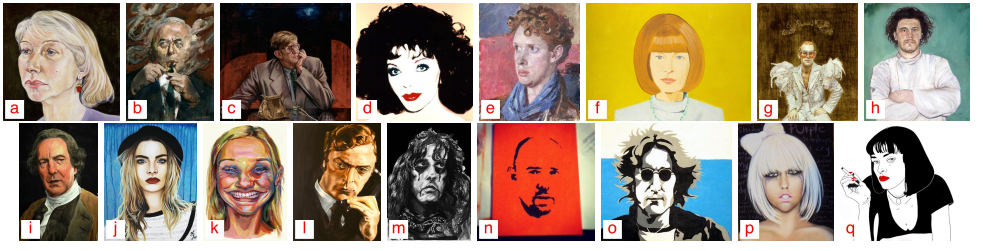

Figure 1: Top row – Paintings from the National Portrait Gallery: (a) Helen Mirren, (b) Harold Wilson, (c) Alan Bennett, (d) Joan Collins, (e) Dylan Thomas, (f) Anna Wintour, (g) Elton John, (h) Marco Pierre-White. Bottom row – Paintings crawled from DeviantArt: (i) Alan Rickman, (j) Cara Delevingne, (k) Cameron Diaz, (l) Michael Caine, (m) Alice Cooper, (n) Karl Pilkington, (o) John Lennon, (p) Lady Gaga, (q) Uma Thurman.

We explore the differences of using shallow [29] vs. deep [28] features for faces for this domain adaptation problem. Furthermore, we study whether retrieval performance can be improved, over using the raw features, by learning either (i) a linear projection on the features using discriminative dimensionality reduction (DDR), or (ii) face-specific classifiers. Section 2 describes the learning methods, section 4 the implementation details, and section 5 assesses the performance. Section 6 considers the inverse problem: how reciprocal is the domain adaptation problem? Given a single painting, can we retrieve photos of that person?

Lastly, in section 7, to return to our original question, we query a large dataset of oil paintings with photos of famous faces to find out if they have any previously unknown doppelgängers, this is further combined with attribute classifiers to retrieve paintings with specific facial attributes such as 'frowning'.

1.1 Related Work

Photos to Paintings. Work on the domain adaptation problem of learning from photos and retrieving paintings has come into being in recent years. Shrivastava *et al.* [35] use an Exemplar SVM [25] to retrieve paintings of specific buildings. Aubry *et al.* [8] improve on this by utilising mid-level discriminative patches, the patches in question demonstrating remarkable invariance between photos and paintings. Subsequently, in [13] we showed that this patch-based method can be extended to object categories in paintings beyond the instance matching of [8]. Others [40, 41] have considered the wider problem of generalising across many depictive styles (e.g. photo, cartoon, painting) by building a depiction invariant graph model.

Features generated using Convolutional Neural Networks (CNN) have shown effectiveness at a variety of tasks [16, 20, 27, 30]. In our earlier work [14], we showed that classifiers

using CNN features learnt on photos are able to correctly recognise object categories in paintings to a high degree of accuracy. We examine here whether this can be extended to facial identities.

Face Identification. There is a vast corpus of work on facial identification (for recent work, see references on the 'Labeled Faces in the Wild' dataset (LFW) website [1]). In many cases [9, 12] the image representation used is highly tailored and face specific. The inspiration for the facial representation in this paper is the work of Simonyan *et al.* [37] who illustrated that a generic Fisher Vector representation [29] performed well on the LFW benchmark [22]. The performance was further improved by discriminatively reducing the dimension of the feature vector by optimising classification loss for positive and negative image pairs, an idea first explored for faces by [21]. An alternative is to optimise a ranking loss, for example over image triplets [11, 19, 34, 39]. Recently, Schroff *et al.* [33] have achieved excellent results on face recognition by learning a CNN using triplet-loss.

2 Learning to improve photo-painting based retrieval

In this section, the methods for using photos to retrieve paintings are described. Assume we have a dataset $\mathcal{D}$ containing paintings of many different people where each painting is represented by a feature vector, y_j. Given a person, the dataset $\mathcal{D}$ is queried using n photos of that person, represented by feature vectors $x_1, x_2 \ldots x_n$.

Three methods are considered: using (i) L2 distance on the original features, (ii) Discriminative Dimensionality Reduction on the photo and painting features, or (iii) by learning classifiers. Both (ii) & (iii) utilise a training set of photos and paintings to learn how to transfer between the two. The details of the features and how the learning methods are implemented are given in section 4.

2.1 L2 Distance

For a given person, each painting in $\mathcal{D}$ is scored according to the mean Euclidean distance between its feature and those of the photos used to query. More formally, given photos $x_1, x_2 \ldots x_n$, the score for each painting y_j is given by $\frac{1}{n}\sum_{i=1}^{n} ||x_i - y_j||_2^2$. The paintings are ranked according to this score, from lowest to highest.

2.2 Discriminative Dimensionality Reduction (DDR)

Here we learn a discriminative linear projection W such that the L2 distance between projected features of a photo and painting, given by $||Wx - Wy||_2^2$, is small if the photo and painting are of the same person and larger by a margin if they are not. There are three reasons for discriminatively learning to reduce the dimension: firstly, it removes redundancy in the feature vectors, allowing them to become smaller, thus more suitable for large-scale retrieval; secondly, it tailors the features to specifically distinguish between faces, which would otherwise be lacking in the case of Fisher Vectors; thirdly, it specifically addresses the domain adaptation between photos and paintings.

The projection is learnt using ranking loss on triplets [34]: given a photo of a person x, a painting of the same person y_+ and a painting of a different person y_-, the projected distance between x and y_+ should be less than that between x and y_- by some margin:

$$||Wx - Wy_+||_2^2 + \alpha < ||Wx - Wy_-||_2^2 \tag{1}$$

Given sets of triplets, W can be learnt by optimising the following cost function which incorporates the constraint (1) softly in a hinge-loss:

$$\underset{W}{\text{argmin}} \sum_{triplets} \max[0, \alpha - (||Wx - Wy_-||_2^2 - ||Wx - Wy_+||_2^2)] \tag{2}$$

This optimisation is carried out using stochastic gradient descent: at each iteration t a triplet (x, y_+, y_-) is considered, and if the constraint (1) is violated, W_t is updated by subtracting the sub-gradient, as:

$$W_{t+1} = W_t - \gamma W_t(x - y_+)(x - y_+)^T + \gamma W_t(x - y_-)(x - y_-)^T \tag{3}$$

where γ is the learning rate. For retrieval, all features are projected by W before L2 distance is calculated, and then paintings are ranked on the mean distance, as above in section 2.1.

2.3 Learning Classifiers

Instead of considering distances, it is possible to learn classifiers using the query photos of each person. As we query with a small number of photos, this is very similar to an Exemplar SVM formulation [25]. Given photos for a person, a linear SVM is learnt that discriminates these query photos from both paintings and photos not containing that person.

3 Data

As described in section 1, retrieval is performed on two distinct datasets containing portraits of people with known identities. Photos of these people are required to query these datasets. In addition to this, the learning methods of section 2 require a training set of both photos and paintings. In this section we describe how the images are sourced (section 3.1), and then how these are used to form the required datasets (section 3.2). A summary of these datasets is provided in table 1.

Dataset	Contents	No. People	Total Images
DEVret	1088 known paintings, 2000 distractor paintings	1,088	3,088
DEVquery	1088 sets of 5 photos to query **DEVret**	1,088	5,440
NPGret	188 known paintings, 2000 distractor paintings	188	2,188
NPGquery	188 sets of 5 photos to query **NPGret**	188	940
Train	248,000 photos and 9,000 paintings for learning	496	257,000

Table 1: The statistics for the datasets used in this paper. 'No. People' refers to the number of known identities among the people present in the dataset. The datasets are described in section 3.2.

3.1 Image Sources

DeviantArt. The website DeviantArt [3] showcases art produced by the public, sorted by various categories (e.g. photography, traditional art, manga, cartoons). Among this work there are many portraits of well known figures, particularly in popular culture. To obtain these portraits, we compiled a list of thousands of famous men and women (people who appeared frequently on IMDB [5]) and crawled DeviantArt using their names as queries. The paintings obtained from DeviantArt are highly prone to variation. Some have been painted and others sketched, many are caricaturistic in nature and lack a photo-realistic quality. The extent of this variety is made clear in the sample paintings provided in figure 1.

National Portrait Gallery. Many of the paintings in London's National Portrait Gallery are publicly available as a subset of the 'Your Paintings' [2] dataset. Some example portraits are shown in figure 1. The portraits are typically quite photo-realistic in nature and are predominantly painted using oil.

3.2 Datasets

We require three types of dataset: (i) query sets, that contain multiple photos of each person; (ii) retrieval sets, that contain paintings of the same persons; and (iii) a training set containing both photos and paintings of people, where the matrix W and classifiers will be learnt from. The query set is used to issue queries for each person, and the performance is measured on the retrieval set. There should be no people in common between the training and other sets. Furthermore, none of the retrieval identities are used to learn the network that produces CNN features.

Retrieval Set – DEVret. A single painting for each of 1,088 people obtained from DeviantArt form a retrieval set. To make the retrieval task more difficult this set is supplemented with 2,000 random portraits from DeviantArt that do not contain any of the people's names in the title.

Retrieval Set – NPGret. A painting for each of 188 people in the National Portrait Gallery is taken to form a retrieval set. The reason this number is not higher is because many people depicted in the National Portrait Gallery lived before the age of photography. These 188 portraits are supplemented with 2,000 random portraits from 'Your Paintings'.

Training Set – TRAIN. The training set consists of multiple paintings per person for each of 496 people from DeviantArt coupled with 500 photos per person from Google Image Search. Some examples of photo-painting pairs with the same identity are given in figure 2. There are 9,000 paintings in total. The distribution of paintings per person is a long tail, this is illustrated in table 2, along with the names of the most prevalent people.

Query Sets. The sets of photos used for querying the retrieval sets, denoted as **DEVquery** and **NPGquery** each contain five photos from Google Image Search per person in their respective sets. The photos have been manually filtered to ensure that they have the correct identities.

Figure 2: Photo-painting pairs that share an identity in the **TRAIN** set. In each case the photo is on the left and the painting is on the right. (a) Tim Minchin, (b) Carey Mulligan, (c) Zooey Deschanel, (d) Dita Von Teese, (e) Hugh Grant.

Name	No. Paintings	Name	No. Paintings
Jared Leto	220	Taylor Swift	184
Clint Eastwood	167	Katy Perry	183
Robert Pattinson	144	Megan Fox	122
Tom Hiddleston	122	Dita Von Teese	89
David Tennant	103	Kate Moss	89
Billie Joe Armstrong	96	Keira Knightly	83
Ian Somerhalder	95	Adriana Lima	76

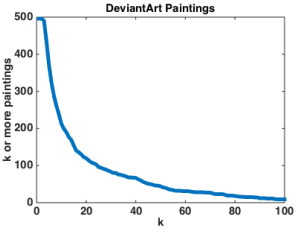

Table 2: Left: the table shows the men and women for which there are the most paintings in **TRAIN**. Right: A plot that shows the number of people for which there are k or more paintings.

4 Implementation Details

Here we describe in detail the feature representations used for the faces as well as the implementation of the methods of section 2.

Face Detection. For a given image, a Deformable Part Model (DPM) [17, 18] trained using the method of [26] is used to detect the location of the face. The detection box is then expanded by 10% in each direction and cropped from the image. The cropped face is then used to compute either a Fisher Vector or CNN feature.

Fisher Vector Representation. For generating improved Fisher Vector [29] features the cropped face is first resized to 150×150 pixels, before the pipeline of [10] is used with the implementation available from the website [9]: RootSIFT [7] features are extracted at multiple scales from each image. These are decorrelated and reduced using PCA to 64-D and augmented with the (x,y) co-ordinates of the extraction location. For each image, the mean and covariance of the distances between its features and each of the 512 centres of a pre-computed Gaussian Mixture Model are recorded and stacked resulting in a 67584-D Fisher Vector. Finally this vector is L2-normalised.

CNN Representation. We use the network of [28]. It is based on the architecture of the VGG Very Deep Model A [36] which achieved state of the art results on the ImageNet Large Scale Recognition Challenge [31]. The network is learnt from scratch using a large dataset of faces of people obtained from Google Image Search. The network is trained using a multi-way classification soft-max loss as described in [36], using the publicly available MatConvNet package [38]. To obtain a feature vector, the cropped face is resized to 224×224 pixels before being passed into the network. The 4096-D output of the penultimate layer (the last fully-connected layer) is then extracted and L2-normalised.

Discriminative Dimensionality Reduction. A PCA projection to 128-D is learnt using the training data. This is used as the W to initialise the optimisation. Triplets are either (i) generated at random offline, or (ii) semi-hard negative triplets [33] are formed online. In the latter case, at each iteration a positive photo-painting pair (x, y_+) is considered with each of n random negative paintings y_- as candidate triplets (we set $n = 100$). Only the candidate for which $(||Wx - Wy_-||_2^2 - ||Wx - Wy_+||_2^2)$ has the lowest positive value is then used. The optimisation is run for 1 million iterations.

Learning Classifiers. For each query, the photos are used as positive examples in an SVM. The negative examples are taken to be all the paintings and photos in the training data. The regularisation parameter C is learnt on a held-out validation set as $C = 1$.

5 Experiments

In this section, the retrieval task is assessed on the two datasets. For each person in the query set (**DEVquery** or **NPGquery**), the photos of that person are used to rank all the paintings in the retrieval set (**DEVret** or **NPGret**) using a given method. The rank held by the correct painting (the one that has the same identity as the query photos) is recorded. Across all people queried, the recall at k for all k is recorded and averaged – this average recall is denoted as Re@k.

The Re@k for various k are given in table 3 for different methods. The corresponding curves are shown in figure 3. A selection of successful retrievals are illustrated in figure 4.

Experiment	Dataset	Re@1	Re@5	Re@10	Re@50	Re@100	Re@1000
FV L2 distance	DEV	4.4	7.4	10.1	17.7	23.2	57.5
FV DDR (i)	DEV	5.9	13.8	18.3	37.0	46.4	91.5
FV DDR (ii)	DEV	7.4	16.3	21.8	40.6	49.4	92.4
FV Classifier	DEV	16.8	25.9	30.1	39.8	46.0	81.3
CNN L2 distance	DEV	26.0	42.2	47.3	63.3	71.4	93.7
FV L2 distance	NPG	4.3	13.3	17.6	25.5	33.0	72.3
FV DDR (i)	NPG	8.5	18.6	23.9	47.3	57.4	95.7
FV DDR (ii)	NPG	7.4	26.6	33.5	54.2	66.0	97.3
FV Classifier	NPG	15.6	24.5	28.7	42.0	49.4	87.2
CNN L2 distance	NPG	36.2	58.5	66.0	80.9	83.0	94.7

Table 3: Percentage Re@k on the retrieval sets for assorted methods and features. DDR refers to Discriminative Dimensionality Reduction, (i) and (ii) refer to the methods of triplet selection given in section 4.

Fisher Vector Learning Results. Both DDR and classification boost the Re@k performance over raw L2 distances for a range of k. This shows that the domain adaptation learning is successful in overcoming the low level statistical and stylistic differences between the photos and paintings. For DDR, Re@k is generally higher for method (ii) (i.e. when semi-hard negative triplets are generated online) as it forces the learning to cope with the most difficult borderline cases, allowing it to distinguish between very similar looking people. Using a classifier (which can learn discriminatively what differentiates a particular identity from others) typically outperforms DDR for low k but is thereafter surpassed. Re@k on

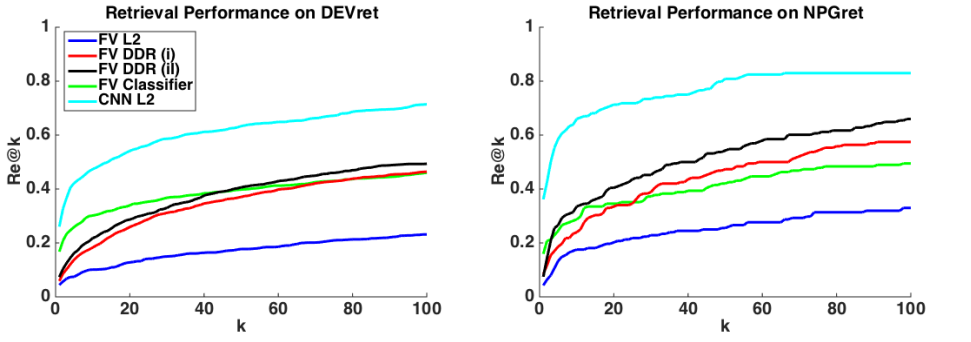

Figure 3: Re@k vs. k plots for DEVret (left) and NPGret (right). The legend on the left plot also applies to the right plot.

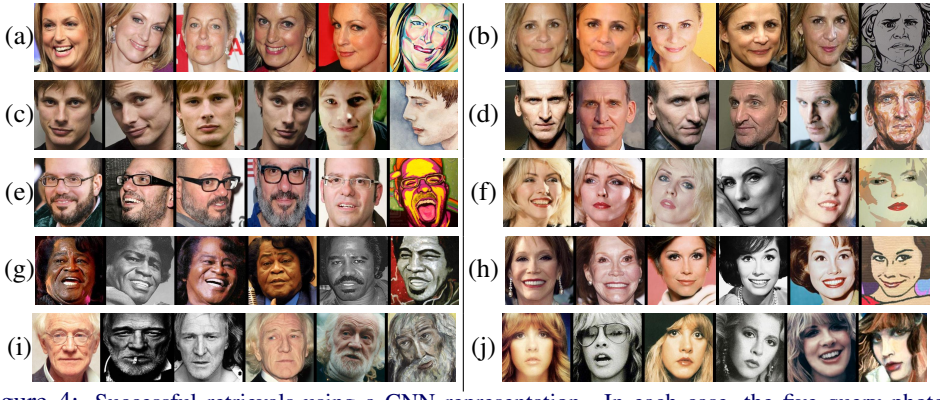

Figure 4: Successful retrievals using a CNN representation. In each case, the five query photos are shown beside the top retrieved painting. (a) Alexandra Wentworth, (b) Amy Sedaris, (c) Bradley James, (d) Christopher Ecclestone, (e) David Cross, (f) Deborah Harry, (g) James Brown, (h) Mary Tyler Moore, (i) Richard Harris, (j) Stevie Nicks.

NPG is generally higher than that on **DEV** for all methods, this is probably because some DeviantArt-style paintings are highly abstract and difficult to retrieve. Interestingly, the DDR matrix performs very well for National Portrait Gallery retrieval, despite having been learnt on DeviantArt style paintings.

CNN Results. The first thing to note is that CNN results always exceed those of Fisher Vectors, even after learning. Interestingly, additional learning for CNN features has negligible effect on performance (this is why further CNN experiments are absent from table 3). The network training has probably already captured the discriminative aspects of a person's face, remarkably to such an extent that the stylistic differences and low level statistics of paintings are of little consequence. The features also demonstrate invariance to pose: notice in figure 4(c) that the side-profile painting of Bradley James has been retrieved using front-profile images. In the case of the Deborah Harry (f) painting where much of the facial outline is missing, the discriminative eyes and lips have been picked out.

6 Retrieving Photos with Paintings

The focus of the majority of this paper has been: starting with photos of a person, retrieve a painting of that person. Here, we instead try to retrieve photos starting with the paintings, to observe how reciprocal the adaptation problem is.

Evaluation. For each of the 1,088 people featured in **DEVret** paintings, 75 photos are crawled from Google Image Search. These are supplemented with distractor photos to form a retrieval set of 97,545 photos. Photos are retrieved from this set using each of the 1,088 **DEVret** paintings as a single query; photos are ranked using the L2 distance between CNN features of the photos and the painting. This proves to be highly successful and some example retrievals are given in figure 5 along with the Average Precisions (AP) of retrieval.

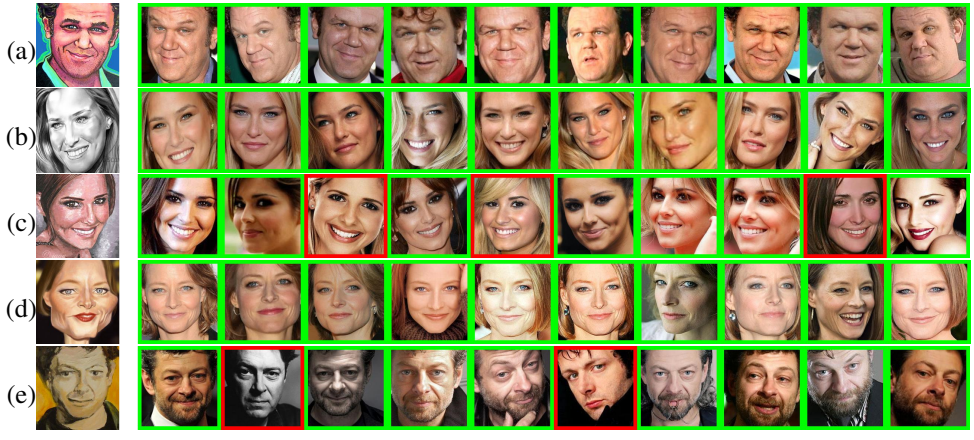

Figure 5: Photo retrieval using a single painting. Each row from left to right shows the painting used for retrieval followed by the 10 highest ranked photos. Correct retrievals have a green border and incorrect retrievals a red one. (a) John C. Reilly AP: 0.90, (b) Bar Refaeli AP: 0.63, (c) Cheryl Fernandez-Versini AP: 0.42, (d) Jodie Foster AP: 0.56, (e) Andy Serkis AP: 0.47

7 Finding Doppelgängers in Art

In this section, we return to our first goal: Given a photo of a person, we would like to retrieve a painting of a very similar looking person. To qualitatively evaluate this problem, we form a set of 40,000 paintings by applying a face detector to the entirety of the 'Your Paintings' [2] dataset and filtering the paintings with the highest classifier scores. This set is queried with photos of famous people known **not** to be present among the paintings, using L2 distances between the CNN features. Some example retrievals are found in figure 6. Notice that the results are uncanny: the portraits and photos have very similar facial features.

Incorporating Attributes. Consider not just being able to find a similar looking painting, but one that also has a given attribute, such as 'frowning'. Motivated by this, we combine this retrieval with attribute classifiers, such that the painting retrieved y satisfies:

$$\underset{y}{\operatorname{argmin}} ||x-y||_2^2 - \lambda w_a y \tag{4}$$

where x is the query photo and w_a is an attribute classifier. λ adjusts the influence of the attribute classifier on retrieval. We demonstrate how retrieval changes with λ in figure 7: as λ increases, so does the extent of the attribute in the retrieved painting.

Implementation Details. Classifiers are learnt for 50 of the attributes listed in [24]: For a given attribute, photos are obtained by crawling Google Image Search with the attribute name as a query. The CNN features extracted from these photos are used as positive examples in a linear SVM with photos of the 49 other attributes as negatives to learn a classifier. These classifiers are then applied to the set of 40,000 paintings.

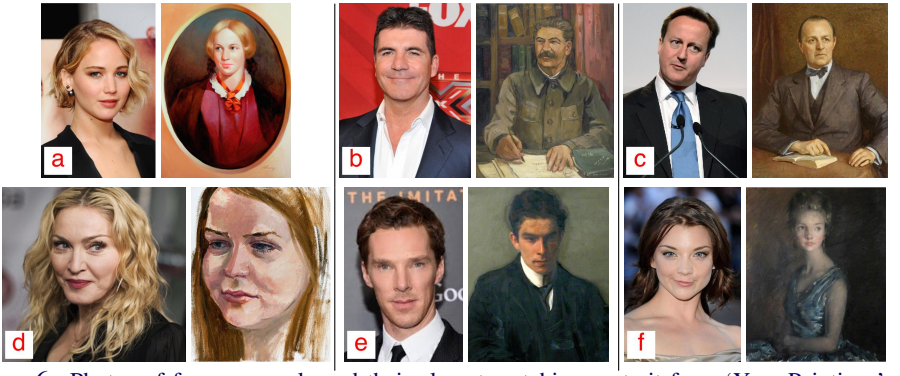

Figure 6: Photos of famous people and their closest matching portrait from 'Your Paintings'. (a) Jennifer Lawrence, (b) Simon Cowell, (c) David Cameron, (d) Madonna, (e) Benedict Cumberbatch, (f) Natalie Dormer.

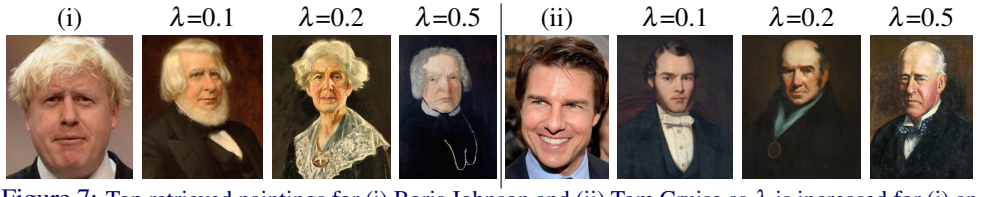

Figure 7: Top retrieved paintings for (i) Boris Johnson and (ii) Tom Cruise as λ is increased for (i) an 'Old face' Classifier and (ii) a 'frowning' Classifier respectively.

8 Conclusion and Future Work

In this paper, we have shown that it is possible to retrieve paintings of people starting with photos of the same person (and vice versa), and that for a Fisher Vector face representation, discriminative learning can significantly increase performance. We have further shown that CNN features produced from a network learnt entirely on photos are able generalise remarkably well to paintings of many different styles. Furthermore, the similarity between these features can be used to find photos and paintings of people that look eerily similar. It would be interesting to explore whether retrieval can be improved by learning a network directly on both photos and paintings.

Acknowledgements. Funding for this research is provided by the EPSRC.

References

[1] LFW Face Database. http://vis-www.cs.umass.edu/lfw/.

[2] BBC – Your Paintings. http://www.bbc.co.uk/arts/yourpaintings/.

[3] DeviantArt. http://www.deviantart.com/.

[4] Encoding methods evaluation toolkit. http://www.robots.ox.ac.uk/~vgg/software/enceval_toolkit/.

[5] Internet movie database. http://www.imdb.com.

[6] ABC News. http://abcnews.go.com/blogs/headlines/2012/11/man-finds-his-doppelganger-in-16th-century-italian-painting/

[7] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, 2012.

[8] M. Aubry, B. Russell, and J. Sivic. Painting-to-3D model alignment via discriminative visual elements. In *ACM Transactions of Graphics*, 2013.

[9] T. Berg and P. N. Belhumeur. Tom-vs-Pete classifiers and identity-preserving alignment for face verification. In *Proc. BMVC*, 2012.

[10] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. BMVC*, 2011.

[11] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research*, 11:1109–1135, 2010.

[12] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High dimensional feature and its efficient compression for face verification. In *Proc. CVPR*, 2013.

[13] E. J. Crowley and A. Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *Proc. BMVC*, 2014.

[14] E. J. Crowley and A. Zisserman. In search of art. In *Workshop on Computer Vision for Art Analysis, ECCV*, 2014.

[15] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.(JAIR)*, 26:101–126, 2006.

[16] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.

[17] P. F. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

[18] P. F. Felzenszwalb, R. B. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2010.

[19] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS*, 2006.

[20] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[21] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *Proc. ICCV*, 2009.

[22] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[23] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.

[24] N. Kumar, A. C. Berg, P. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. ICCV*, 2009.

[25] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *Proc. ICCV*, 2011.

[26] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistle. In *ECCV*, 2014.

[27] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Proc. CVPR*, 2014.

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. BMVC*, 2015.

[29] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

[30] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *DeepVision Workshop, CVPR*, 2014.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, S. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and F.F. Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[32] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010.

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, 2015.

[34] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2004.

[35] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics*, 2011.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015.

[37] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In *Proc. BMVC*, 2013.

[38] A. Vedaldi and K. Lenc. MatConvNet – Convolutional Neural Networks for MATLAB. *CoRR*, abs/1412.4564, 2014.

[39] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[40] Q. Wu and P. Hall. Modelling visual objects invariant to depictive style. In *Proc. BMVC*, 2013.

[41] Q. Wu, H. Cai, and P. Hall. Learning graphs to model visual objects across different depictive styles. In *Proc. ECCV*, 2014.